5/31/17

"Coming together is a beginning keeping together is progress working together is success."
-Henry Ford

HW: "Regression" homework section
Test 3 Wednesday 6/7

AIM: What is Regression?

Warm Up:

2. $\lim_{x\to -3} \frac{x^2-9}{2x^2+7x+3}$

3. $\lim_{h\to 0} \frac{(2+h)^3-8}{h}$

4. $\lim_{x\to 1} \frac{x^8-1}{x^5-x}$

1. Scientists randomly select ten groups from a population of men over 50 years old. They calculate the mean weights of each of these groups. The variability between these means can be best attributed to

(1) measurement variability (3) induced variability

(2) natural variability (4) sampling variability

Variability is mainly being introduced by the random sampling being done.

(4)

2. Max and Daniel are measuring the amount of time it takes for a ball to roll down a ramp at different heights. For each trial, both Max and Daniel take turns rolling the ball and working the stop watch. They do this in order to quantify which of the following sources of variability?

(1) measurement variability (3) induced variability

(2) natural variability (4) sampling variability

Since Max and Daniel are humans, they will time the ball rolling down the ramp slightly differently. Thus they are trying to quantify measurement variability.

(1)

3. Which of the following scenarios would be an attempt to quantify induced variability?

Induced variability means subjects are assigned to different treatment groups.

(1) a phone survey of political preferences during election season

(2) multiple random samples of products from an assembly line to check for defects

(3) random assignment of people to a control group and a group taking a drug to lower cholesterol

(4) recording the variability in the measurement of a soil sample's weight by the same machine

(3)

4. Which of the following research questions would involve collecting data through a survey?

(1) Watching people exit a grocery store to see the percent who use reusable bags.

(2) Assigning people to two groups to see the effect of a particular amount of sleep.

(3) Calling people on the telephone to see if they will be voting in the upcoming election.

(4) Dropping salt cubes into two different liquids to determine which dissolves faster.

(3)

5. In which of the following cases would an observational study be necessary as compared to an experimental study?

In an observational study, the conductors of the study do not get to assign the subjects to the different groups.

(1) The study of how increased nutrient levels affect plant growth.

(2) The study of how educational levels affect median household income.

(3) The study of how a vaccine affects the percent of mice that get a particular disease.

(4) The study of how noise level affects the sleep patterns of volunteers in a sleep study.

(2)

1. Which of the following formulas, written in summation notation, would represent the mean of the data set $\{x_1, x_2, ..., x_n\}$? Explain your choice.

(1) $\sum_{i=1}^{n} x_i$ (3) $n\sum_{i=1}^{n} x_i$

(2) $\frac{1}{n}\sum_{i=1}^{n} x_i^2$ (4) $\frac{1}{n}\sum_{i=1}^{n} x_i$

To calculate the mean, you simply add all of the data values, i.e. $\sum_{i=1}^{n} x_i$, and then divide by how many there are, i.e. $\frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$

(4)

2. The standard deviation of a population characteristics measures

(1) The difference between the maximum and minimum values.
(2) The difference between the third quartile and first quartile values.
(3) The average distance a data value is away from the mean.
(4) The average distance a data value is away from the median.

(3)

3. The interquartile range of the data set $\{4, 7, 10, 13, 18, 22, 30\}$ is

(1) 15 (3) 7

(2) 18 (4) 10

$Q_1 = 7$ and $Q_3 = 22$
IQR = 22 – 7 = 15

(1)

APPLICATIONS

4. If 348 freshmen out of 622 have cell phones, then the population proportion, p, for freshmen cell phone ownership is

(1) 0.56 (3) 0.72

(2) 0.35 (4) 0.44

$p = \frac{348}{622} = 0.5594... \approx 0.56$

(1)

5. If a population has 824 subjects, then about how many would have characteristics in the upper quartile?

(1) 412 (3) 368

(2) 280 (4) 206

$\frac{1}{4} \cdot 824 = 206$

(4)

6. A school is tracking its freshmen attendance for the first marking period. Shown below is a table summarizing their findings for the 284 members of the freshmen class.

Days Absent (x_i)	Number of Students (f_i)
0	158
1	64
2	18
3	22
4	4
5	7
6	8
9	2
13	1

(a) Find the mean and median number of days absent. Round your mean to the nearest tenth.

mean = $\bar{x}$ = 1.042... ≈ 1.0 days
median = 0 days

(b) What is the population standard deviation for this data set? Round to the nearest tenth.

σ_x = 1.763... ≈ 1.8 days

(c) What proportion of the population that has an absenteeism greater than 4 days?

Count the number of students with 5 days or more absent:
7 + 8 + 2 + 1 = 18
$$p = \frac{18}{284} = 0.063... \approx .06 \text{ or } 6\%$$

7. The heights of the 15 players on the Arlington boys' varsity basketball team are given below in inches.

66, 67, 68, 68, 70, 72, 72, 73, 74, 75, 75, 75, 76, 77, 79

(a) Find the mean and standard deviation of this data set. Use the population standard deviation. Round both to the nearest tenth.

mean = $\bar{x}$ = 72.466... ≈ 72.5 in
std dev = σ_x = 3.7923... ≈ 3.8 in

(b) Determine the proportion of the population that falls within one standard deviation and within two standard deviations of the mean. State your values in decimal form.

One standard deviation from the mean:

72.5 − 3.8 = 68.7
72.5 + 3.8 = 76.3
$$p = \frac{9}{15} = 0.60$$

Two standard deviations from the mean:

72.5 − 2(3.8) = 64.9
72.5 + 2(3.8) = 80.1
$$p = \frac{15}{15} = 1.00$$

(c) Use the random number table for this lesson to pick a random sample of five players from this list. Do this by picking a random two digit column along the page. Scan down the column until you have picked 5 random integers that fall from 1 to 15. Write down your sample and calculate its mean.

I generated a random sample and got the numbers:

6, 9, 10, 11, and 13 which gave me a data set of:

72, 74, 75, 75, 76

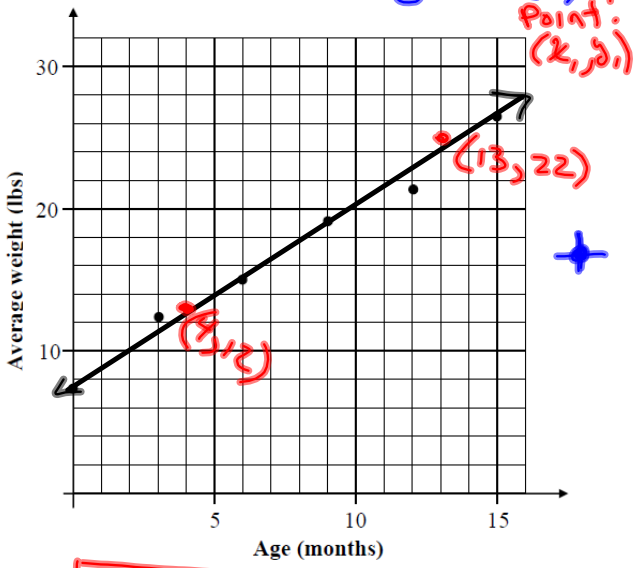
mean = $\bar{x}$ = 74.4 inches

***Exercise* #1:** A pediatrician would like to determine the relationship between infant female weights versus age. The pediatrician studies 100 newborn girls and finds their average weight at the end of 3 month intervals. The data is shown below and graphed on the scatter plot.

Age (months)	0	3	6	9	12	15
Average Weight (pounds)	7.2	12.2	15.1	19.4	21.5	26.3

$y-y_1=m(x-x_1)$ & $y=mx+b$ slope (m) y-int (b)

(a) Using a ruler, draw a line that you think best fits this data. As a general guideline, try to draw it such that there are as many data points above the line as below it.



(b) By picking two points that are on the line (not necessarily data points), determine the equation of your best fit line. Round your coefficients to the nearest *tenth*.

(4, 12) (13, 22)

$m=\frac{22-12}{13-4}=\frac{10}{9}$

$y-12=\frac{10}{9}(x-4)$

$y-12=\frac{10}{9}x-\frac{40}{9}$ → $y=1.1x+7.6$

(c) Using the linear regression command on your calculator, find the equation of the best fit line for this data. Round all **linear parameters** to the nearest *tenth*.

$y=1.2x+7.8$

(d) Use your calculator to determine the **linear correlation coefficient.** Round to the nearest *thousandth*. How can you interpret this value in terms of the variation in weight due to age?

$r=.995$

There is a high correlation between age and weight b/c r is very close to 1

***Exercise* #2:** Using the equation that your calculator produced in Exercise #1, predict the weight of a baby girl after 10 months. Round your answer to the nearest tenth of a pound.

plug in 10 for x

$y = 1.2x + 7.8$

$y = 1.2(10) + 7.8 =$ 19.8 pounds

The use of a model to predict outputs when the input is within the range of the known data is called **interpolation**. Interpolation tends to be fairly accurate.

inside of what we have

***Exercise* #3:** Using the equation that your calculator produced in Exercise #1, predict the weight of a baby girl after 2 years. Round your answer to the nearest tenth of a pound.

24 months

$y = 1.2(24) + 7.8 =$ 36.6 pounds

The use of a model to predict outputs when the input is outside of the range of the known input data is called **extrapolation**. Models are most helpful when they can be used to extrapolate, but tend to be less accurate.

OUTSIDE DATA

***Exercise* #4:** Biologists are trying to create a least-squares regression equation (another name for best fit line) relating the length of steelhead salmon to their weight. Seven salmon were measured and weighed with the data given below.

Length (inches)	22	24	28	34	39	42	48
Weight (pounds)	3.43	4.46	7.08	14.21	22.19	31.22	35.67

(a) Determine the least-squares regression equation, in the form $y = ax + b$, for this data. Round all coefficients to the nearest *hundredth*.

$y = ax + b$

$a = 1.32554$

$b = -27.9849$

$r^2 = .968$

$r = .983$

$y = 1.33x - 27.98$

(b) Using your equation from part (a), determine the expected weight of a salmon that is 30 inches long.

$y = 1.33(30) - 27.98$

$=$ 11.92 pounds

interpolation.

(c) Using your equation from part (a), determine the expected weight of a salmon that is 52 inches long.

41.18 pounds

extrapolation

(d) In which part, (b) or (c), did you use interpolation and in which part did you use extrapolation? Explain.

Just as we fit data with a linear model we can also fit with all sorts of other mathematical models, depending on the context of the situation. In this lesson we will examine **exponential regression** and **sinusoidal regression**. You could be asked to run a quadratic regression, logarithmic regression, power regression, …. The process is similar for each and all are found in Stat Calc menu. Exponential regression is review from Common Core Algebra I, so we will start with that.

***Exercise* #5:** The population of Jamestown has been recorded for selected years since 2000. The table below gives these populations.

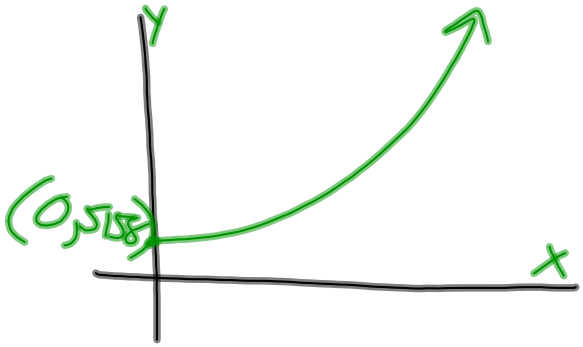
x= 2 4 5 7 9

Year	2002	2004	2005	2007	2009
Population	5564	6121	6300	6812	7422

(a) Using your calculator, determine a best fit exponential equation, of the form $y = a \cdot b^x$, where x represents the number of years since 2000 and y represents the population. Round a to the nearest integer and b to the nearest *thousandth*.

$y = 5158(1.041)^x$

(b) Sketch a graph of the exponential function for the years 2000 to 2050. Label your window and your y-intercept.



(c) By what percent does your exponential model predict the population is increasing per year? Explain.

growth factor is 1.041

4.1%

(d) Algebraically determine the number of years, to the nearest year, for the population to reach 20 thousand.

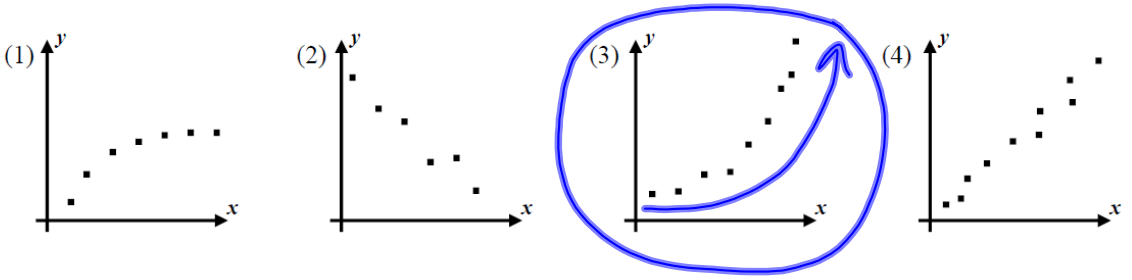
$\frac{20000}{5158} = \frac{5158(1.041)^x}{5158}$

$\frac{20000}{5158} = (1.041)^x$

$\log_{1.041} \frac{20000}{5158} = x$

$x \approx 34$ years

***Exercise* #6:** Which of the following scatter plots would be best fit with an exponential equation?



Sinusoidal, or trigonometric, regression is much more complicated than either linear or exponential. It should be used in situations that appear **periodic** in nature.

***Exercise* #7:** The temperature of a chemical reaction changes during the reaction. The temperature was measured every two minutes and the data is shown in the table below.

Time (min)	0	2	4	6	8	10	12	14	16	18	20
Temp $(^\circ C)$	35.7	38.9	41.6	42.3	40.8	38.4	36.1	34.2	35.9	39.1	41

(a) Why does it seem like this data might be periodic? Create a quick scatter plot using your calculator to verify.

(b) Use your calculator to do a sine regression in the form $y = a\sin(bx+c)+d$. Round all parameters to the nearest tenth. Graph along with your data to informally assess the fit of the curve. When prompted use 16 iterations always.

(c) According to this model, what is the range in temperatures the chemical reaction will include?

(d) According to this model, what is the time it takes for the reaction to complete one full cycle?

***Exercise* #8:** The maximum amount of daylight that hits a spot on Earth is a function of the day of the year. Taking $x = 0$ to be January 1st, daylight, in hours, was measured for 12 different days. The measurement was the number of possible hours of sun from sunrise to sunset.

Day	0	34	68	98	118	134	171	203	274	321	346
Daylight Hours	9.0	9.9	11.5	13.1	14.0	14.6	15.2	14.8	13.1	11.5	9.5

(a) What is the natural period of this data set?

(b) Use your calculator with the period from (a) to find an equation of the form $y = a\sin(bx+c)+d$ that fits this data, then examine the graph of the equation on the scatter plot. How good is the fit?

(c) What is the maximum amount of daylight hours predicted by the model? Show your calculation.